Análise de similaridade dos tuítes sobre a guerra na Ucrânia

Trabalho de Machine Learning

Tópicos

Introdução e Objetivo

Metodologia

Pré Processamento

Mineração

Pós Processamento e Conclusão

Introdução e Objetivo

- Analisaremos a semelhança entre tuítes escritos sobre a Guerra na Ucrânia utilizando a métrica TF-IDF para a seleção de atributos
- Tuítes postados por usuários com mais de 5 000 000 seguidores no Twitter por um período de mais de 90 dias
- Formato do trabalho: aplicação, contendo as fases de pré-processamento, mineração e pós-processamento

Metodologia

DataSet Original

- Conjunto de DataSets contém 34.21M tuítes
- Os DataSets contém 17 colunas (principais: *"text", "username", "followers"* e *"language"*)

```
Unnamed: 0                                                    332446
userid                                                    1115874631
username                                                  CGTNOfficial
acctdesc          CGTN is an international media organization. I...
location                                                 Beijing, China
following                                                        74
followers                                                  13383261
totaltweets                                                  216669
usercreatedts                                2013-01-24 03:18:59.000000
tweetid                                          1509300280248217600
tweetcreatedts                               2022-03-30 22:43:22.000000
retweetcount                                                     1
text              #Ukraine and #Russia had hold the 5th round of...
hashtags          [{'text': 'Ukraine', 'indices': [0, 8]}, {'tex...
language                                                        en
coordinates                                                    NaN
favorite_count                                                   7
extractedts                                  2022-03-30 22:54:18.659938
Name: 5632, dtype: object
```

Bibliotecas utilizadas

- Numpy - biblioteca de matemática computacional
- Pandas - implementação do DataFrame
- NLTK - métodos e funções de processamento de linguagem natural
- SciKit Learn - possui algoritmos de classificação, regressão, clusterização, etc, e métodos para a criação de atributos, como o TF-IDF;
- SciPy - contém métodos de ML, como linkage e dendrogram para a implementação e criação de clusters hierárquicos
- Mlxtend (machine learning extensions) - ferramentas úteis para as atividades diárias de ciência de dados, sendo utilizada para a aplicação do algoritmo de regras de associação.

Pré Processamento

- Os datasets foram filtrados baseando-se nos valores das colunas “language” e “followers”
- Criação de um novo DataSet
- Para cada usuário, separamos os tuítes e aplicamos uma *tokenização* junto a um filtro com as *stop words* seguido de uma *stemmização* dos tokens
- Cálculo da métrica TF-IDF *(term frequency–inverse document frequency)* para cada conjunto de tuítes de cada usuário.

- Lista das palavras mais relevantes conforme as frequências TF-IDF, que serão utilizadas para análise de similaridade
- Filtro de palavras baseadas naquelas que aparecem para uma porcentagem *p* de usuários
- O dataset final, para p = 15% é composto por 114 linhas e 157 colunas

	#	$	%	&	//t.co/	@	across	aid	amid	amp	...	video	wa	watch	way	week	woman	work	world	would	year
IndiaToday	0.111044	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
htTweets	0.096142	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
the_hindu	0.119334	0.0	0.0	0.0	0.000000	0.322511	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TimesNow	0.245121	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ndtv	0.203021	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
TEDTalks	0.208514	0.0	0.0	0.0	0.000000	0.208514	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.208514	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SkySports	0.277350	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BBCNews	0.824163	0.0	0.0	0.0	0.000000	0.137361	0.0	0.0	0.0	0.0	...	0.137361	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
paugasol	0.200000	0.0	0.2	0.0	0.000000	0.200000	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0
BenStiller	0.426401	0.0	0.0	0.0	0.213201	0.000000	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0

114 rows × 157 columns

Mineração

Clusterização Hierárquica

- Método de clusterização baseado em alguma métrica e método de distância distinto.
- Métrica do cosseno: medir a semelhança entre objetos de um mesmo conjunto de dados
- Dois métodos diferentes de clusterização: duas abordagens que serão utilizadas para calcular a distância a partir da métrica utilizada
  - single - menor distância entre dois clusters
  - complete - maior distância entre dois clusters
- Criação de dois dendrogramas, um para cada método escolhido

Regras de Associação e Apriori

- Apriori:
  - Consiste na identificação dos diferentes conjuntos (sets) de atributos que frequentemente aparecem no Dataset
- Regras de associação:
  - Calcula diferentes métricas de associação para a descoberta de relações entre os sets de atributos adquiridos pelo Apriori
  - Uma associação (relação) $A \rightarrow B$ é composta de um set antecedente (A) e um set consequente (B)

As métricas calculadas são:

- *support* (range: [0,1]) - frequência/importância de uma relação num Dataset
- *confidence* (range: [0,1]) - probabilidade de um set consequente acontecer, dado um set antecedente
- *lift* (range: [0,∞]) - medida de quanto o antecedente e o consequente apareceriam juntos para caso eles fossem independentes

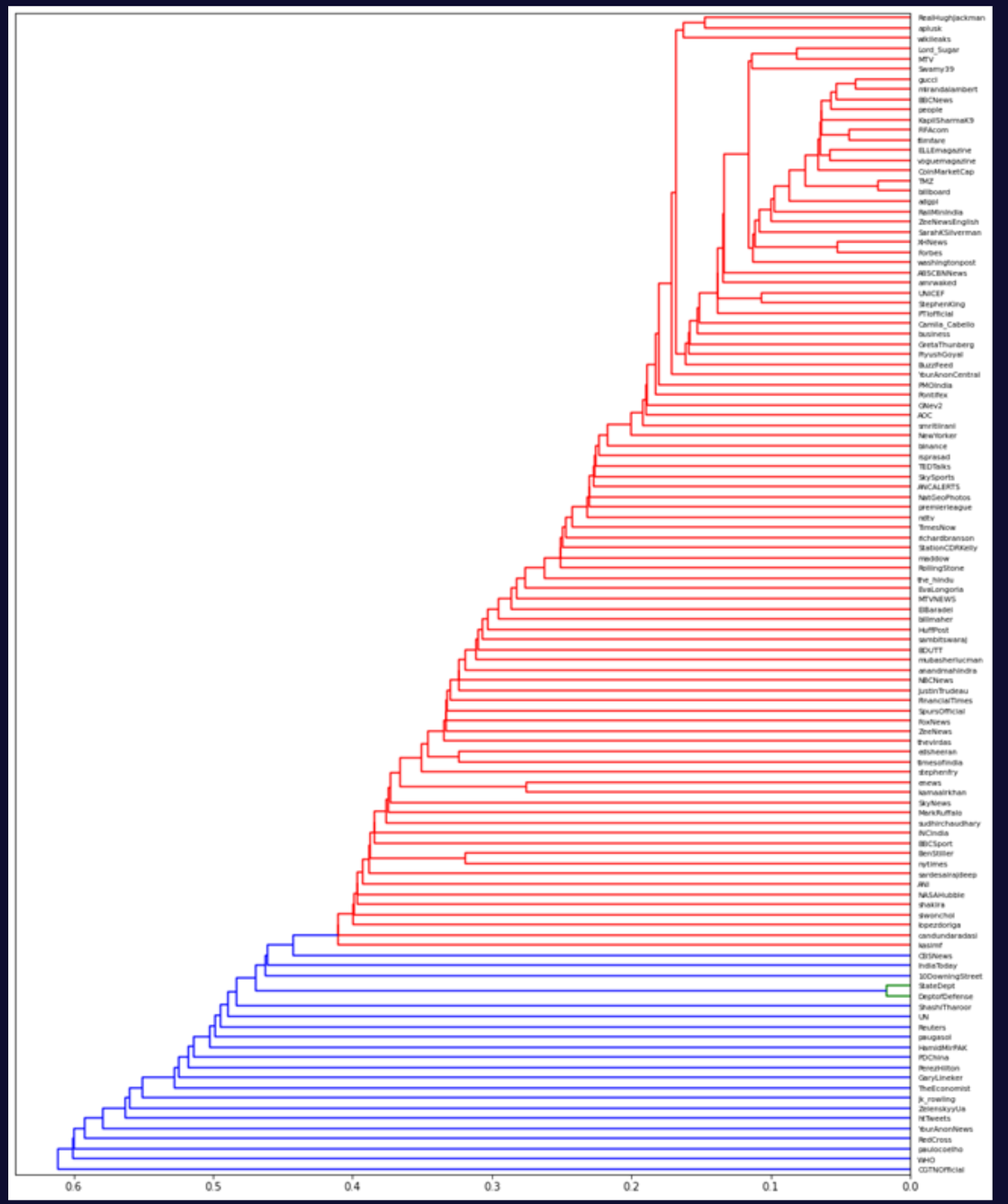
- *leverage* (range: [-1,1]) - cálculo da diferença entre a frequência observada do antecedente e consequente aparecerem juntos e a frequência que seria esperada se ambos fossem independentes
- *conviction* (range: [0, ∞]) - métrica de dependência do consequente perante o antecedente.

Para este trabalho, foram selecionados sets com support acima de 50% e regras de associação com confidence acima de 60%, resultando em 43 regras dadas como interessantes pelo algoritmo.

Pós Processamento e Conclusão

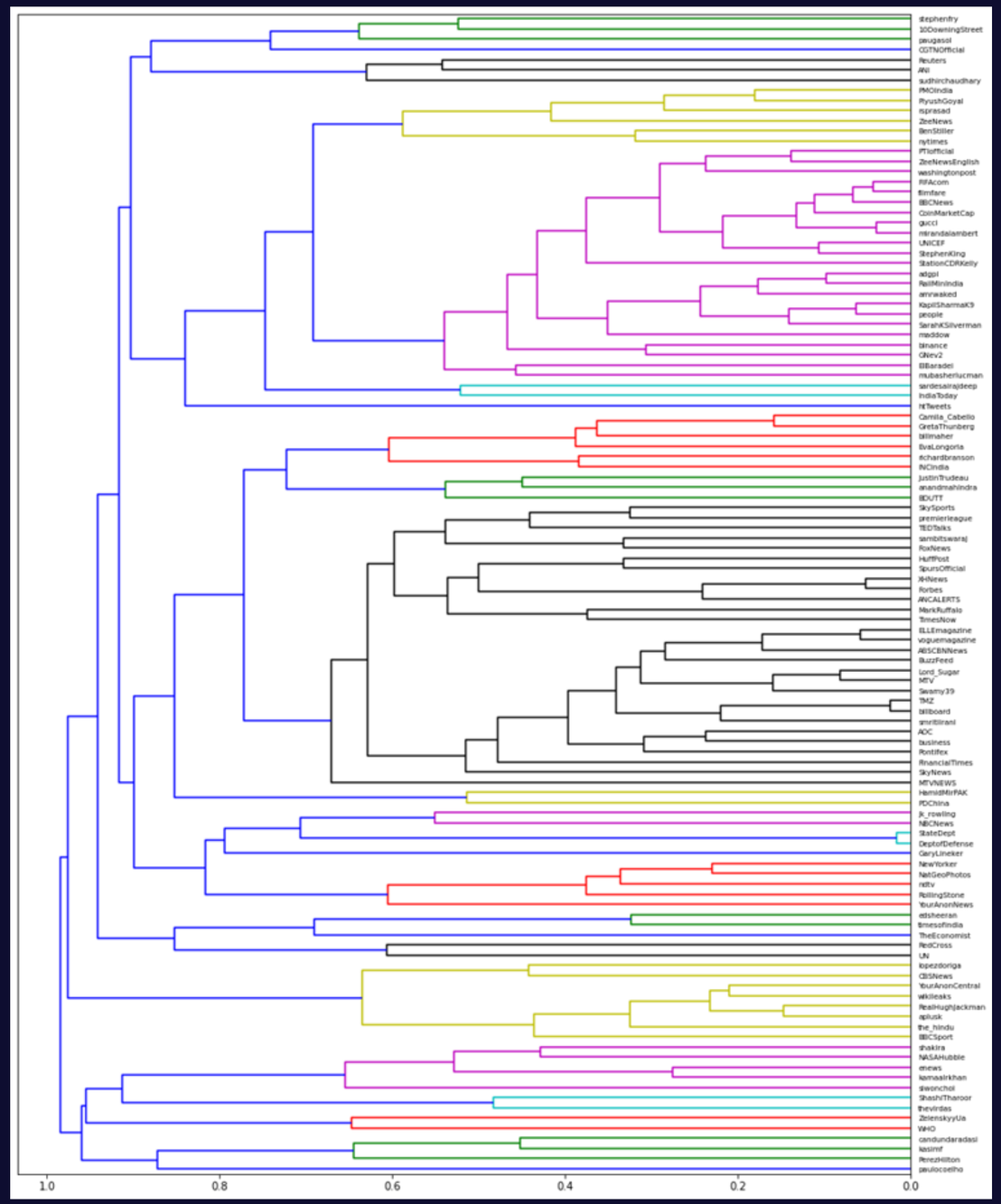
Dendograma gerado a partir do método de distância *single*

- Formato em "escada"
- Os objetos clusterizados não possuem relação e, portanto, não há características comuns para agrupar a maioria dos objetos em grupos

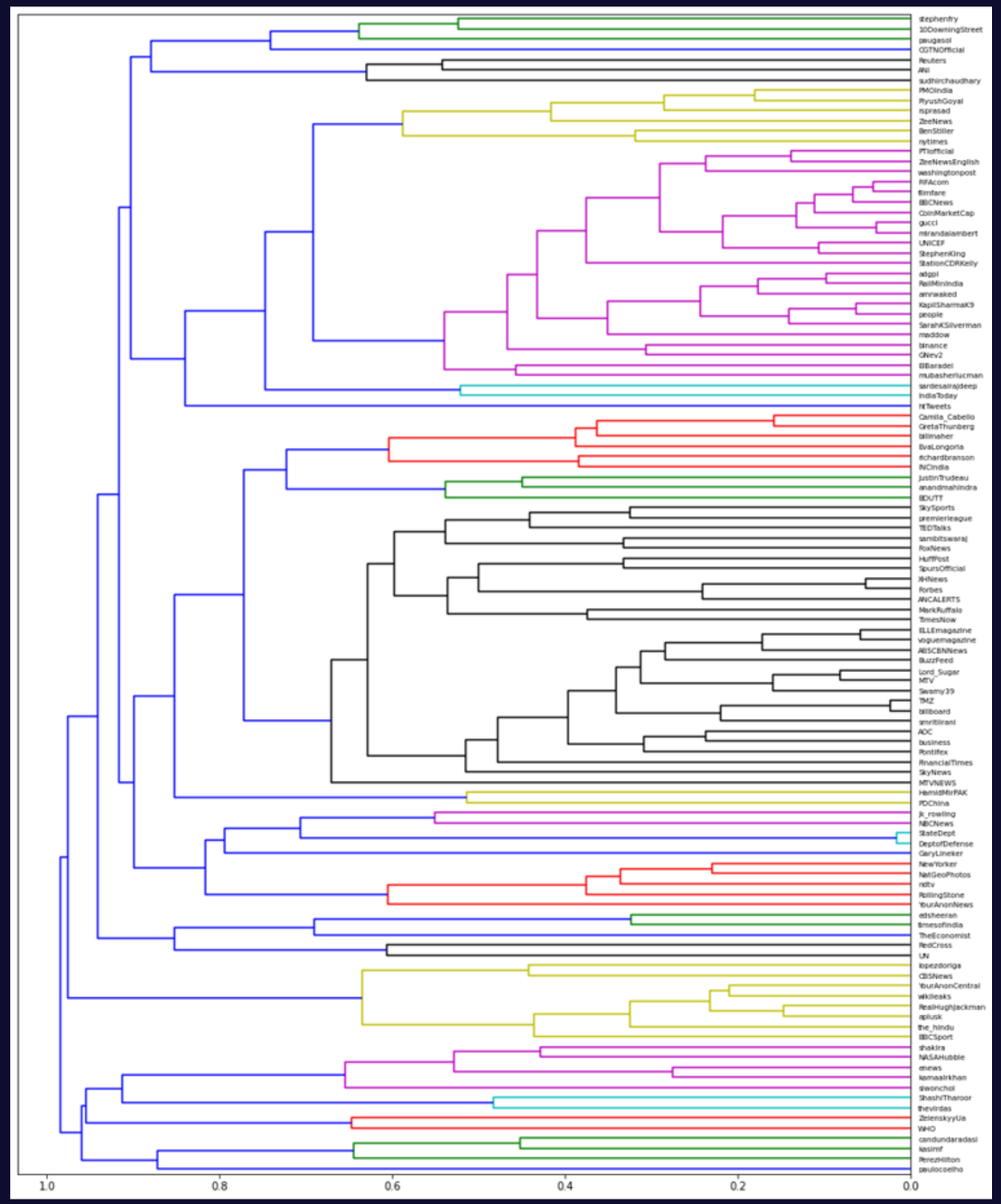


Dendograma gerado a partir do método de distância *complete*

- É possível perceber o agrupamento de alguns poucos grupos interessantes (ex: Departamento de Defesa dos EUA junto do Departamento do Estado dos EUA ou WikiLeaks e YourAnonCentral)



- Porém, além dessas e algumas outras poucas relações, não há uma consistência dos grupos perante tipos de conta como de jornais e canais de notícia, como BBC ou CNN; atores e cantores, como Mark Ruffalo (Hulk); e grandes sites de entretenimento ou cosméticos, como a marca de roupas Gucci.



Regras de associação

- Os valores de *lift* e *leverage* indicam uma independência dos sets
- Maior valor de *lift* encontrado é 1.08, enquanto o menor é 1.00
  - valores iguais a 1 indicam independência
- Maior valor de *leverage* encontrado é 0.04, enquanto o menor é 0.00
  - valores iguais a 0 indicam independência
- O conjunto de palavras que foram selecionadas pelo apriori e pertencem a diferentes sets de interesse é muito pequeno
  - *#*, *ukraine*, *http* e *numericalValue*
- Dos 157 atributos, apenas as palavras #, ukraine, http e numericalValue, aparecem nas relações

- As palavras são, em sua maior parte, independentes entre si e os textos dos diversos usuários não seguem um padrão de uso delas

- Algo importante a se notar é o valor de support alto para as palavras *#*, *http* e *ukraine*, indicando que os tuítes dos usuários com o maior número de seguidores consequentemente utilizam de hashtags, links e das palavras chaves dos tópicos em alta para atrair leitores para seus tuítes, ferramentas de conhecimento comum para o aumento do alcance de uma conta no Twitter

Obrigado!

Gustavo Fernandes Carneiro de Castro – 11369684
Nayara Kellen Peralta – 11345235